Linear Discriminant Analysis (LDA): Supervised Dimensionality Reduction and Classification

Ahmed BADI
ahmedbadi905@gmail.com
linkedin.com/in/badi-ahmed

January 2026

Abstract

Linear Discriminant Analysis (LDA) is a classic technique used both for classification and supervised dimensionality reduction. Unlike PCA, which is unsupervised and focuses only on capturing variance, LDA explicitly uses class labels to find projections that best separate different classes. It does this by maximizing the ratio of between-class variance to within-class variance. In this article, we develop the intuition behind LDA, derive its mathematical formulation via scatter matrices and the Fisher criterion, and show how it can be used to reduce dimensionality while preserving discriminative information. We also discuss assumptions, practical considerations, and comparisons with PCA and other methods.

Keywords: Linear Discriminant Analysis, Fisher Criterion, Supervised Dimensionality Reduction, Scatter Matrices, Classification.

1 Introduction

In many classification problems, data live in a high-dimensional space, but only a few directions are actually useful for separating classes. LDA provides a way to find these directions using label information. [1], [2]

While PCA seeks directions of maximum variance without considering labels, LDA seeks directions that maximize separation between class means while minimizing spread within each class. This makes LDA particularly well-suited for dimensionality reduction in supervised settings, where the goal is to improve classification performance and interpretability. [3], [4]

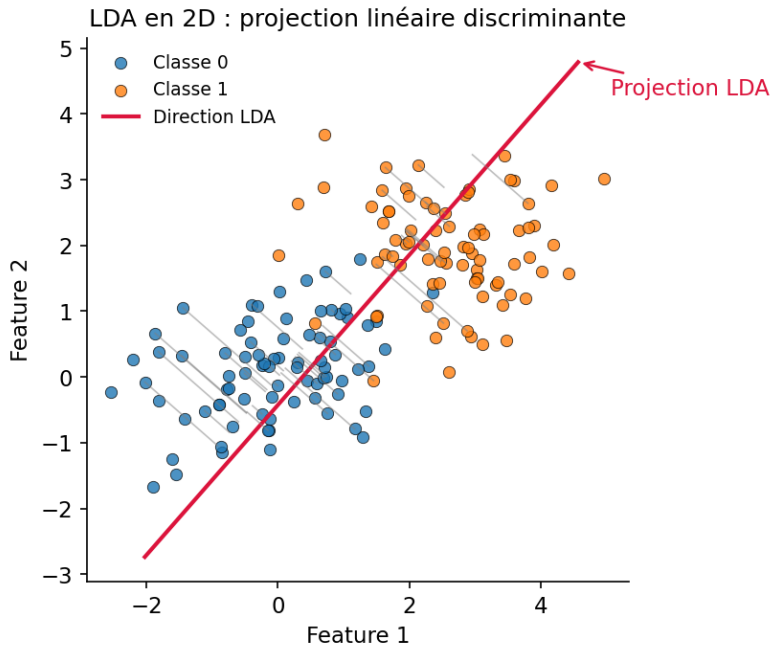

Figure 1: LDA finds a projection (red line) that maximizes separation between class means and minimizes within-class variance.

2 Problem Setup

Suppose we have a labeled dataset:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N},$$

where $\mathbf{x}_i \in \mathbb{R}^p$ are feature vectors and $y_i \in \{1, \ldots, K\}$ are class labels. We assume:

- There are K classes.
- Class k has N_k samples.

LDA aims to find a linear transformation

$$\mathbf{z} = W^\top \mathbf{x}, \quad W \in \mathbb{R}^{p \times d},$$

where $d < p$, such that in the projected space, classes are as well separated as possible. For classification, d is typically at most $K - 1$. [1], [4]

3 Scatter Matrices

To measure separation, LDA defines two scatter matrices: **within-class scatter** and **between-class scatter**. [3], [5]

3.1 Class Means

Let

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} \mathbf{x}_i$$

be the mean of class k, and

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

be the global mean.

3.2 Within-class Scatter Matrix

The within-class scatter matrix $S_W \in \mathbb{R}^{p \times p}$ measures how samples spread around their class means:

$$S_W = \sum_{k=1}^{K} \sum_{i:y_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top.$$

Each term $(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$ contributes to the scatter inside class k. [1], [5]

3.3 Between-class Scatter Matrix

The between-class scatter matrix S_B measures how far class means are from the global mean:

$$S_B = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top.$$

If class means are far apart, S_B will have large values. [3], [5]

4 Fisher Criterion for Two Classes

For intuition, consider two classes ($K = 2$) and a projection onto a single direction $\mathbf{w} \in \mathbb{R}^p$:

$$z_i = \mathbf{w}^\top \mathbf{x}_i.$$

We want:

- The projected class means to be far apart.
- The projected within-class variance to be small.

Let projected class means be:

$$m_1 = \mathbf{w}^\top \boldsymbol{\mu}_1, \quad m_2 = \mathbf{w}^\top \boldsymbol{\mu}_2.$$

Projected within-class variances are:

$$s_1^2 = \sum_{i:y_i=1} (\mathbf{w}^\top \mathbf{x}_i - m_1)^2, \quad s_2^2 = \sum_{i:y_i=2} (\mathbf{w}^\top \mathbf{x}_i - m_2)^2.$$

Fisher proposed maximizing the criterion: [5], [6]

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}.$$

This ratio increases when the means are far apart and the within-class variance is small.

4.1 Matrix Form

It can be shown that for two classes, Fisher's criterion can be written as:

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}.$$

The optimal $\mathbf{w}$ (up to scaling) that maximizes $J(\mathbf{w})$ is:

$$\mathbf{w}^\star \propto S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

So the best projection direction is given by the inverse within-class scatter multiplied by the difference of class means. [1], [5]

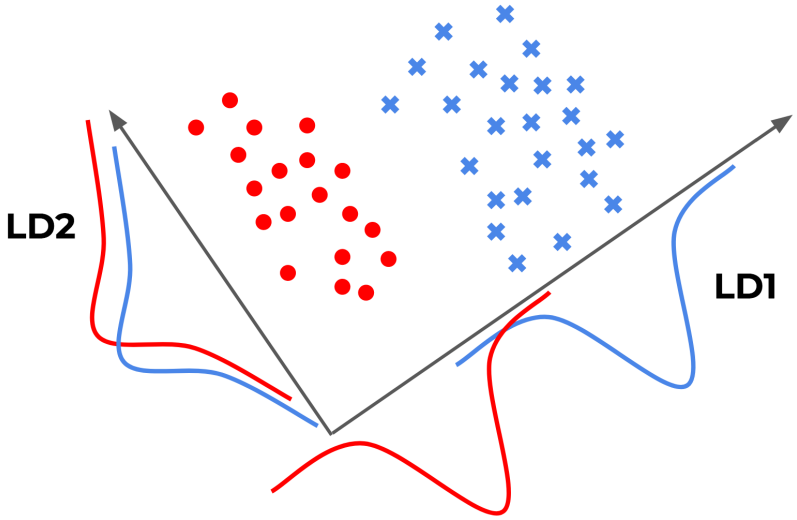

Figure 2: Fisher's criterion maximizes separation of projected means while minimizing within-class spread.

5 Multi-class LDA and Generalized Eigenproblem

For $K > 2$ classes and a projection to dimension d (typically $d \leq K - 1$), LDA generalizes Fisher's criterion to:

$$J(W) = \frac{\det(W^\top S_B W)}{\det(W^\top S_W W)},$$

or using trace:

$$J(W) = \mathrm{Tr}\left((W^\top S_W W)^{-1} W^\top S_B W\right).$$

Maximizing this leads to the generalized eigenvalue problem: [1], [5]

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}.$$

The columns of W are the eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$ corresponding to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$. [5], [7]

The low-dimensional features are then:

$$\mathbf{z}_i = W^\top \mathbf{x}_i \in \mathbb{R}^d.$$

6 LDA as Supervised Dimensionality Reduction

LDA can be seen as a supervised dimensionality reduction method: [4], [8]

- It projects data onto a subspace of dimension at most $K - 1$.
- It chooses the subspace to maximize class separability, measured by the Fisher criterion.

In scikit-learn, for example, calling `LinearDiscriminantAnalysis(n_components=d)` and then `fit_transform` computes W and returns the projected data. [4]

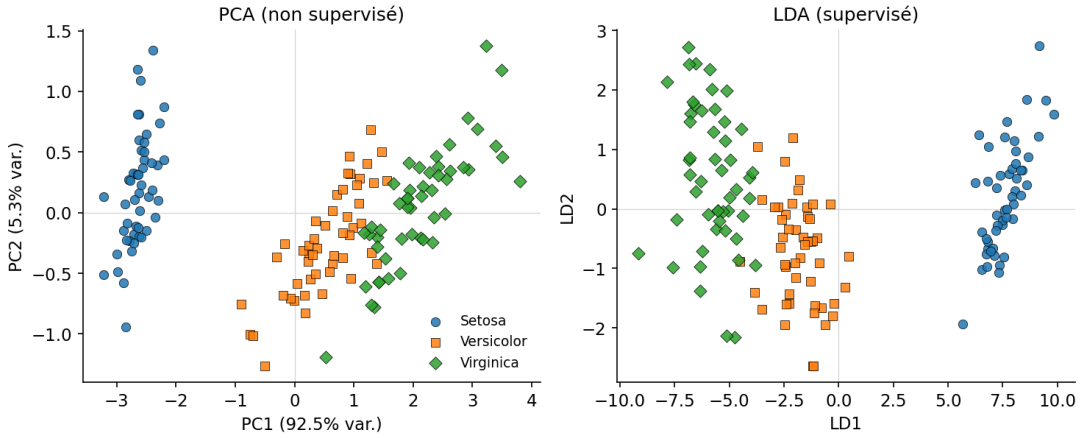

Figure 3: Comparison of PCA and LDA on the Iris dataset: LDA uses labels and tends to separate classes better.

7 LDA for Classification

Originally, LDA is also a linear classifier under Gaussian assumptions. [1], [2]

7.1 Gaussian Class-Conditional Densities

LDA assumes:

- Each class k has a multivariate Gaussian distribution:

$$p(\mathbf{x} \mid y = k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma),$$

  with a common covariance matrix Σ across classes.

- Prior probabilities $P(y = k)$ may be known or estimated from data.

Under these assumptions, the posterior $P(y = k \mid \mathbf{x})$ leads to linear decision boundaries: [9]

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \log P(y = k).$$

The predicted class is:

$$\hat{y} = \arg\max_k \delta_k(\mathbf{x}).$$

7.2 Connection to Fisher LDA

The discriminant directions obtained from the generative Gaussian model are related to the Fisher directions from the scatter-matrix viewpoint; both aim to separate classes based on means and a shared covariance structure. [5], [9]

8 Practical Considerations

8.1 Assumptions

LDA works best when: [3], [10]

- Classes are approximately Gaussian.
- Covariance matrices of classes are similar (homoscedasticity).
- Classes are roughly linearly separable in some projection.

Violations of these assumptions may reduce performance; in such cases, quadratic discriminant analysis (QDA) or nonlinear methods might be better. [1], [4]

8.2 Choice of Dimensionality

The maximum number of useful LDA components is $K - 1$, where K is the number of classes. [4], [8]

- For $K = 2$, LDA reduces to a single discriminant direction (1D).
- For multi-class problems, we can visualize data in 2D using the first two linear discriminants.

8.3 Comparison with PCA

Key differences: [8], [11]

- PCA is unsupervised and ignores labels; LDA is supervised and uses labels.
- PCA maximizes total variance; LDA maximizes class separability.
- PCA can have up to p components; LDA has at most $K - 1$ components.

9 Applications

LDA is used in many areas: [1], [10]

- **Preprocessing** for classification: project data to a low-dimensional space with better class separation.
- **Pattern recognition**: face recognition, handwriting recognition, speech recognition.
- **Bioinformatics**: supervised dimensionality reduction for gene expression or single-cell data. [12]

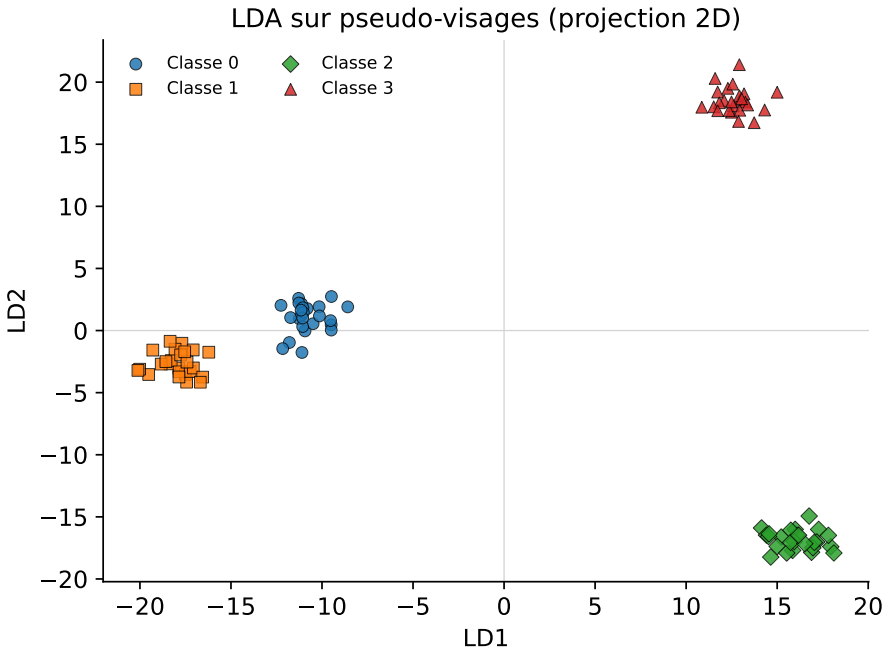

Figure 4: LDA applied to face images: projected features improve class separation between individuals.

10 Conclusion

Linear Discriminant Analysis is a powerful supervised tool that:

- Performs dimensionality reduction by projecting data into a low-dimensional subspace (up to $K - 1$ dimensions).
- Maximizes class separability by using the Fisher criterion on scatter matrices.
- Provides linear decision boundaries under Gaussian assumptions with shared covariance.

While LDA relies on assumptions and linear projections, it often outperforms PCA for classification tasks because it explicitly uses label information. In practice, LDA is a useful baseline for supervised dimensionality reduction and classification, especially when the number of features is high and the number of samples is moderate. [1], [8]

References

[1] Wikipedia contributors, *Linear discriminant analysis*, `https://en.wikipedia.org/wiki/Linear_discriminant_analysis`, 2024.

[2] IBM, *What is linear discriminant analysis?* `https://www.ibm.com/think/topics/linear-discriminant-analysis`, 2023.

[3] Applied AI Course, *Linear discriminant analysis (lda) in machine learning*, `https://www.appliedaicourse.com/blog/linear-discriminant-analysis-in-machine-learning/`, 2024.

[4] scikit-learn developers, *Linear and quadratic discriminant analysis*, `https://scikit-learn.org/stable/modules/lda_qda.html`, 2024.

[5] F. Fleuret et al., *Fisher and kernel fisher discriminant analysis: Tutorial*, http://nvayatis.perso.math.cnrs.fr/FilesCHPS/tutorial-FDA.pdf, 2010.

[6] T. S. Souza, *An illustrative introduction to fisher's linear discriminant*, https://sthalles.github.io/fisher-linear-discriminant/, 2018.

[7] D. Li, *Fisher linear discriminant analysis*, https://www.khoury.northeastern.edu/home/vip/teach/MLcourse/5_features_dimensions/lecture_notes/LDA/LDA.pdf, 2014.

[8] Encord, *Top 12 dimensionality reduction techniques for machine learning*, https://encord.com/blog/dimentionality-reduction-techniques-machine-learning/, 2025.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] DataCamp, *Linear discriminant analysis: Beyond dimension reduction*, https://www.datacamp.com/tutorial/linear-discriminant-analysis, 2025.

[11] S. Saha, *Lda is more effective than pca for dimensionality reduction in classification datasets*, https://towardsdatascience.com/lda-is-highly-effective-than-pca-for-dimensionality-reduction-in-classification-datasets-4489eade9f5e, 2021.

[12] F. J. H. Heras and G. G. de Polavieja, "Supervised dimensionality reduction by a linear discriminant analysis on pre-trained cnn features," *arXiv preprint arXiv:2006.12127*, 2020.